DOCUMENT RESUME

ED 221 547 TM 820 540

AUTHOR Choppin, Bruce
TITLE Criterion-Referencing of Performance by Latent-Trait Scaling.
PUB DATE Oct 81
NOTE 12p.; Paper presented at the Annual Meeting of the Military Testing Association (October 26-30, 1981).

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Academic Achievement; Achievement Tests; *Criterion Referenced Tests; Diagnostic Tests; *Individual Testing; *Latent Trait Theory; Scaling; Student Evaluation; *Test Interpretation; Test Validity

ABSTRACT

Coventional true score approaches to criterion-referenced tests are not applicable when tests are used for classification. Criterion-referenced testing is now used frequently for educational decision making. The assumption is made that these instruments are to be considered as tests designed to classify into "mastery" and "non-mastery" groups, rather than to measure. There is no precise methodology in use to establish cut-off scores beyond which mastery can be assumed to have occurred. Latent trait modeling of test performance can enable straightforward criterion-referenced interpretations of test scores. A future expansion in the use of testing to diagnose the learning problems of individual students is foreseen. Individual diagnostic testing requires criterion-referencing of test performance, and of each item response. Criterion-referencing with latent trait models will become more useful as individual diagnostic testing becomes more prevalent. (Author/DWH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED221547

CRITERION-REFERENCING OF PERFORMANCE BY LATENT-TRAIT SCALING

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.
Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

by
Bruce Choppin
Center for the Study of Evaluation
University of California, Los Angeles

ABSTRACT

Conventional true-score approaches to criterion-referenced tests break down when such tests are used for classificatory purposes. Latent trait modeling can help in this situation, but it is appropriate to question whether this classificatory use of criterion-referenced tests is what is needed. Latent trait modeling of test performance can enable straightforward criterion-referenced interpretations of test scores. As the use of testing for individual diagnosis increases, criterion-referencing with latent trait models is likely to become even more attractive.

TM 820 540

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

B H. Choppin

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Criterion-Referencing of Performance by Latent-Trait Scaling

Over the last 15 years the two major developments in the field of educational measurement have been criterion-referenced testing and the development of latent-trait models. Each topic now commands a substantial proportion of the available space in professional journals and each has benefitted from the close attention of some of the most creative minds and clearest thinkers in our field. Yet, remarkably little of this thinking has been directed towards bringing these two developments together, or into positions from which they could support each other. (A notable exception to this is van der Linden's article in the current issue of the Review of Educational Research to which I shall refer later.)

Criterion-referenced measurement was originally envisaged as a system which permitted the interpretation of test performance in terms of some external criterion. The test score would be used to relate the individual's performance to a meaningful scale of tasks rather than merely conveying information about the individual's standing vis-a-vis other persons who had taken the same test. Performances from the field of athletics, and examinations such as the driving test and measures of typing proficiency were often cited as examples.

In recent years, however, this conception of criterion-referencing has been steadily replaced by another, better suited for use in decision making situations. Educational decisions (in common with many others) frequently depended on which of two categories or

classifications contained a particular object or event. In education once the concept of mastery learning had become fashionable, the classification was often according to whether or not a mastery state had been achieved (Meskauskas, 1976). If achieved, then the instructional program could progress; if not, then some remedial work was indicated. This situation was modeled with idealized domains of items all of which were focused on this particular discrimination between masters and non-masters. One psychometric model on which a considerable amount of work has been published makes the intrinsically satisfying assumption that all the items comprising a domain are of equivalent difficulty, and that mastery of the domain was to be equated with the ability to respond correctly to all the items which comprised it. Other approaches, which allowed for item difficulty variation within the domain, sought to estimate the proportion of the domain which individual students had mastered. Of course these models are not realistically portrayals of actual educational measurement situations. Domains are rarely so well defined. Items are related in varying degrees to the variable to be measured and contain differing amounts of irrelevant and contaminating content. They are not all of equal difficulty. Neither, in many situations, are they selected randomly from the domain for inclusion in tests. Yet the models should not be rejected for these reasons. All models necessarily simplify in order to represent the crucial features of a complex situation, and should not be expected to be a true reflection of reality. Indeed it is precisely this simplification of a real situation that give them a

a potential value.

The attempts to extend classical item analysis procedures to derive indices such as reliability, validity, and discrimination for criterion-referenced tests are carefully considered by van der Linden in the article to which reference has already been made (van der Linden, 1981). He points out that, particularly in the area of test validity, serious objections can be made to all the methods that are in common use. These objections include:

(a) the effect that variations in item difficulty have on the pattern of responses in the pretest-posttest situation which is often used for determining criterion-referenced test validity;

(b) the lack of quality control for the intermediate instructional experiences in the same pretest-posttest design;

(c) the dependence of many of the models (and the parameters they contain) on the size of the variance of performance among the sample of people who provide the validating data.

At issue, of course, are the characteristics of the test instrument at and immediately around the cut-off point. Further, it has been argued (e.g., Glaser & Nitko, 1971) that criterion-referenced validity of a test score requires that the score itself has an external referent, i.e., that it be interpretable in terms of specific levels of behavior in a defined domain. Conventional statistical analyses of criterion-referenced test data do not provide this.

In his review, van der Linden takes on the problem of mastery classification by considering a criterion-referenced test's measurement characteristics: validity, reliability and discrimination at the cut-off point. In this context, he argues for the use of an item information function derived from a latent trait model, and then he demonstrates the application of this approach to criterion-referenced test construction by identifying items that do not discriminate well at the cut-off point.

This approach, employing the item information function, works in theory with all the latent trait models, but for the models with two or three item parameters there may be great difficulties of interpretation. Although such models order a set of items uniquely in terms of their difficulty parameters, this ordering is not in general the same as the ordering of the probabilities for a correct response to each item by any particular student, and in particular by a student at or near the cut-off point. These complex models are proving valuable for detailed study of test-taking behavior under varying conditions, but they do not provide a convenient basis for measuring students, nor for making dichotomous classifications. For these operations we need sets of items that fit one parameter models where such can be developed (Choppin, 1981).

I wish to take issue with only one aspect of van der Linden's review which is implicit throughout his whole treatment, and made explicit in his final paragraph. This is the notion that since criterion-referenced testing is now almost always used for decision making, it is appropriate to treat criterion-referenced tests as instruments designed to classify (into "mastery" and "non-mastery" groups)

rather than to measure. Unfortunately we have no precise methodology in general use for establishing the cut-off point beyond which mastery can be confidently assumed to have occurred. Methods such as the Nedelsky procedure for organizing expert judgment in an objective fashion mask the fact that the necessary bridge between test item response and learning theory has yet to be constructed (Glaser, 1981). I recall from the very early days of the Mastery Learning Movement a time when 80% or more of the students scoring 80% or more on the test was the necessary and sufficient condition for the teacher to infer "mastery" of the material, and to move on to the next unit. As far as I know, there was no empirical evidence to support the use of these percentages. I do remember that rather more than 10 years ago Ben Bloom asked me how many questions you had to ask a student to be sure that he had mastered any single objective. My reply went something along the lines, "For typical test items, you should be looking for a score of 3 out of 3, or if you want to allow the possibility of an accidental error then ask for a score of at least 4 out of 5". I hope that that was not the beginning of the 80% criterion.

To clarify the point, let me turn to my favorite everyday latent-trait, temperature, and the practical realization of its model for measurement - the thermometer. The thermometer is a measuring instrument calibrated so that we can give consistent external meaning to a range of different temperature observations. Where we need only to classify temperatures into two classes around a defined cut-off point, we substitute for the thermometer a degenerate form of it, the thermostat. Thermostats are effective, but they have a

very limited range of applications - and their sensitivity at the cut-off temperature is perhaps their most important feature. The validity, reliability and discrimination (sensitivity) of the thermostat is established empirically by repetitious use. If we were able to perform similar experiments in education to validate cut-off points for mastery learning we might be able to discard measurement. Then when we had established where the cut-off point was, it might be appropriate to use criterion-referenced tests that operated like thermostats.

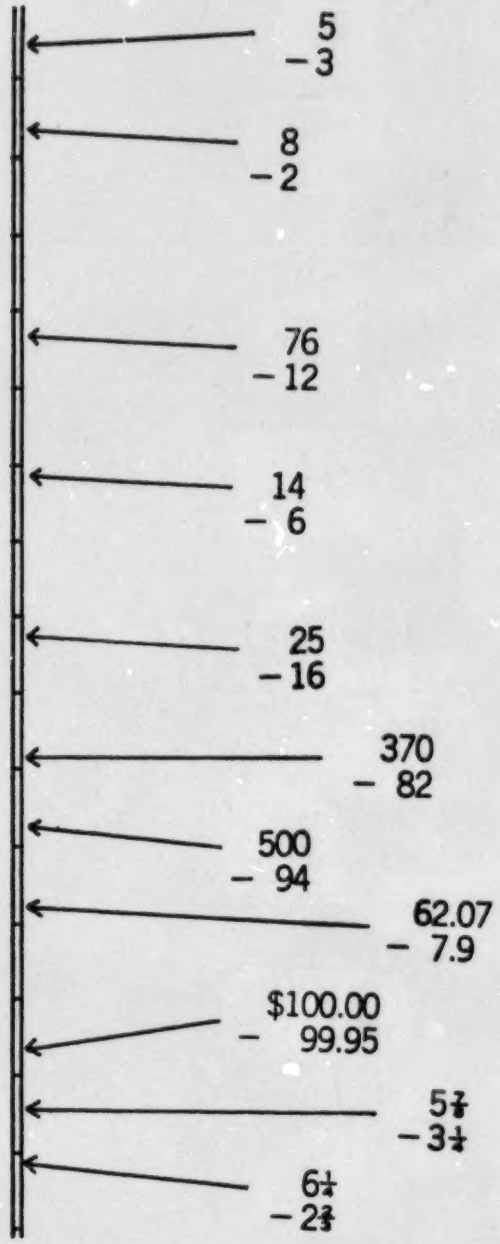
Until that time, we need to use tests as measuring instruments and to calibrate them just as we calibrate general purpose thermometers. We need them to explore the possible effects of using alternative cut-off points on clearly defined scales. Further, these alternative cut-off points need to be defined in terms of some external reference such as sample behaviors or tasks. Latent trait models, since they relate achievement level and item difficulty directly on the same scale, can give such a reference. Two examples which I would bring to your attention are scales for measuring achievement level in arithmetic subtraction (taken from the Keymath Diagnostic Test, Connolly, Nachtman & Pritchett, 1971) and "operations" from the Mathematics Profile Series (Cornish & Wines, 1975). In each, the location of an individual's achievement on this scale gives an immediate visual indication of what material has been mastered, and what remains to be learned.

Another implementation of this approach is to be found in the "KIDMAP" reports on student achievement used (among other places)

MAPS Scale of Difficulty and Ability	Small Numbers	Large Numbers	Pronumerals
25	$3+4=4+\triangle$ $6\times 1=\triangle$		
30	$7\times 8=8\times\triangle$		
35	$15\div 3=\triangle\div 3$	$123+456=456+\triangle$ $44\times 125=125\times\triangle$ $(23\times 24)\times 25=\triangle\times(24\times 25)$ $864\div 432=\triangle\div 432$	$p+q=q+\triangle$ $r\times s=s\times\triangle$
40	$5-2=\triangle-2$ $(3\times 2)\times 5=\triangle\times(2\times 5)$ $9+1=\triangle$ $5+0=\triangle$ $(5+4)+6=\triangle+(4+6)$ $8\times 0=\triangle$	$984\times 1=\triangle$ $876+0=\triangle$ $654-543=\triangle-543$ $578+1=\triangle$ $(89+67)+33=\triangle+(67+33)$	$h\div j=\triangle\div j$ $(u\times v)\times w=\triangle\times(v\times w)$ $m-y=\triangle-y$ $(h+j)+k=\triangle+(j+k)$ $g+0=\triangle$
45	$(8-4)+4=\triangle$	$(987-321)+321=\triangle$	
50	$15\div 5=30\div\triangle$ $(12\div 2)\times 2=\triangle$ $8\div 4=\triangle\div 8$ $7-4=\triangle-7$ $4+5=(4+6)+(5+\triangle)$ $(24\div 6)\div 2=\triangle\div(6\div 2)$	$769\times 0=\triangle$ $(625\div 25)\times 25=\triangle$ $654-543=\triangle-654$ $240\div 15=480\div\triangle$	$y\times 1=\triangle$ $w\times 0=\triangle$ $v+1=\triangle$
55	$(7\times 2)-(3\times 2)=(7+\triangle)\times 2$	$123+456=(123+789)+(456+\triangle)$ $468\div 234=\triangle\div 468$ $(72\times 25)-(60\times 25)=(72+\triangle)\times 25$	$(j-k)+k=\triangle$ $s\div t=rs\div\triangle$ $p+q=(p+r)+(q+\triangle)$ $(m\div g)\times g=\triangle$
60	$(12-6)-4=\triangle-(6-4)$	$(900\div 30)\div 10=\triangle\div(30\div 10)$	$mp-np=(m+\triangle)\times p$
65	$(40\div 8)\times 4=(40\times 4)\div(\triangle\times 4)$	$(89-56)-21=\triangle-(56-21)$	$l-e=\triangle-l$ $k\div p=\triangle\div k$
70		$(72\div 36)\times 9=(72\times 9)\div(\triangle\times 9)$	$(y\div x)\times w=(y\times w)\div(\triangle\times w)$
75			$(u-v)-t=\triangle-(v-t)$ $(p\div q)\div r=\triangle\div(q\div r)$

OPERATIONS SCALE

(Mathematics Profile Series)
Cornish & Wines, 1977



SUBTRACTION SCALE

(Keymath Diagnostic Arithmetic Test ; Connolly, Nachtman & Pritchett, 1971)

within the Los Angeles County schools, in which the curriculum, expressed in terms of objectives, is mapped out on a single sheet of paper, and the student's achievement level is indicated by his location vis-a-vis the objectives (Wright, Mead & Ludlow, 1980).

It seems probable that the future will see a considerable expansion in the use of testing to diagnose the learning problems of individual students and trainees. Much of our testing technology has grown out of group testing situations whether for program evaluation or for the ranking of testees. The context for individual diagnostic testing will be very different. Such testing necessarily requires criterion-referencing of test performance, and indeed of each item response. Latent trait modeling becomes especially attractive in this context because of its efficient integration of diagnostic evidence from different items. We have some of the techniques for doing this, but much more development work remains to be done.

REFERENCES

Choppin, B.H. The use of latent-trait models in the measurement of cognitive abilities and skills. In D. Spearritt (Ed.) The improvement of measurement in education and psychology. Australian Council for Educational Research, 1981.

Connolly, A.J., Nachtman, W. & Pritchett, E.M. Keymath Diagnostic Arithmetic Test. American Guidance Service, 1971.

Cornish, G. & Wines, R. Mathematics Profile Series: Operations Test. Australian Council for Educational Research, 1977.

Glaser, R. The future of testing: A research agenda for cognitive psychology and psychometrics. Technical Reprint No. 3. Univ. of Pittsburgh, LRDC, 1981

Glaser, R. & Nitko, A.J. Measurement in learning and instruction. In R.L. Thorndike (Ed.) Educational measurement, American Council in Education, 1971.

Meskauskas, J.A. Evaluation models for criterion-referenced testing. Review of Educational Research, 1976, 46, 133-158.

van der Linden, W.J. A latent trait look at pretest-posttest validation of criterion-referenced test items. Review of Educational Research, 1981, 51, 379-402.

van der Linden, W.J. Decision models for use with criterion-referenced tests. Applied Psychological Measurement, 1980, 4, 469-492.

Wright, B.D., Mead, R.J. & Ludlow, L.H. KIDMAP: Person-by-item interaction mapping. Research Memorandum No. 29, University of Chicago: MESA Press, 1980.